KENTUCKY TECHNICAL ADVISORY COMMITTEE (KTAC)
Oct. 21, 2021
10:30 a.m. to 5:30 p.m. ET
Virtual Meeting and Conference Room 517
300 Sower Blvd., 5th Floor
Frankfort, KY

KTAC MEMBERS PRESENT: Elena Diaz-Bilello, Pete Goldschmidt, Corrine Huggins-Manley, Suzanne Lane, Phoebe Winter

KDE MEMBERS PRESENT: Karen Dodd, Michael Hackworth, Kevin Hill, Helen Jones, Jennifer Larkins, Felicia Nu'Man, Kevin O'Hair, Mike Prater, Ben Riley, Rhonda Sims, Jennifer Stafford, John Wickizer

KDE GUESTS PRESENT: Bill Auty (EdMeasure, KDE psychometrician); Brian Gong (Center for Assessment, facilitator); Marc Johnson, Eric Moyer, Tim O'Neil, Stanley Rabinowitz, Adrian Riveria, Brad Ungurait, Scott Wilson (Pearson); Emily Dickinson, Art Thacker (HumRRO)

The meeting began at 10:30 a.m. ET.

Agenda Item: Welcome and Legal Requirements

Presenter: Rhonda Sims, KDE Associate Commissioner, Office of Assessment and Accountability, and Felicia Nu'Man, KDE Staff Attorney, Office of Legal Services

Summary of Discussion: Sims welcomed members of KTAC, KDE staff and KDE guests, and provided an overview of the meeting agenda.

Nu'Man presented background of the statutory authorization for KTAC and associated responsibilities, including complying with open meetings requirements.

Agenda Item: Background of Kentucky's Assessment Program

Presenter: Rhonda Sims, Associate Commissioner, KDE Office of Assessment and Accountability; Kevin Hill, Division Director, KDE Office of Assessment and Accountability; and Jennifer Stafford, Division Director, KDE Office of Assessment and Accountability

Summary of Discussion: Sims, Hill and Stafford provided an overview of Kentucky's assessment program, including what is assessed, in which grades and what is new for 2022. In addition, a brief background of Kentucky's accountability system was shared. The focus and purpose of the overview was to help KTAC members understand the context and provide advice regarding the upcoming 2022 assessments. In addition, Kentucky's innovative assessment and

accountability initiatives were mentioned, which would be discussed later in the meeting. KTAC members thanked KDE staff for the informative presentations and did not have any questions.

Agenda Item: Test Design for 2022 Reading and Math Assessments: Reporting, Test Blueprints, Scaling and Equating

Presenter: Rhonda Sims, Associate Commissioner, KDE Office of Assessment and Accountability; Bill Auty, KDE Psychometrician

Summary of Discussion: Sims introduced the technical staff working on Kentucky's 2022 assessments in reading and mathematics. The technical staff provided background and solicited KTAC members' advice on issues of test design, test blueprints, scaling, equating and reporting. These issues are inter-related, and decisions regarding one aspect have implications for the others.

One key challenge is that Kentucky seeks a shorter summative assessment that still provides reliable and valid scores for individual students, but which also would sample the full domain more completely and provide more information for schools and districts. The proposed solution is to design a test that has a core of common items for all students in a grade/content area in reading or mathematics, and that also includes a set of items that could differ by student. This is referred to as "matrix-sampled" items. KDE and contractor staff presented options being considered and KTAC members asked clarifying questions and provided advice and suggestions.

KTAC members recommended that the purpose and intended use of the assessment results be clearly specified, since those will drive the design of the assessment. They noted in particular that the emphasis was first on providing reliable and valid information about individual students for summative accountability purposes, and secondarily for group scores (i.e., at the school or district level) through aggregation of individual student performance, derivative scores (such as growth) and through matrix-sampling, which could inform curriculum planning and program evaluation.

The results of the summative assessment should supplement other assessments, such as interim and formative assessments to inform instructional decisions regarding students through the year, both because those other types of assessments can be more focused in terms of what is assessed in relation to what has been taught, and because results can be provided closer in time to instruction during the year.

KTAC members recommended that the test plans be specified in detail, including: main scores and subscores; distribution of content (e.g., math content or math practices); item response type; item score points; cognitive complexity; and for the reading assessment, text characteristics (such as complexity, length, reading level, etc.). In addition, including what would be common and what would be matrix-sampled for operational reporting, and field testing. The specifications should be designed to support the intended claims, interpretations and uses. Test plans should reflect planned changes over administrations, including providing for field testing, equating and addressing the full depth and breadth of needed evidence.

KTAC members recommended that an item/test development plan be developed in detail that would aim the development of new assessment items enough to create new test forms, taking into account the need for equating, field testing, retirement of items for security and public release of items.

Also discussed was several possible options for reporting, focusing on utility and dimensionality. The proposed matrix-sampled item design that would support group scores at the school and district levels draws on additional evidence than the individual student scores. KTAC members talked about how the matrix-sampled group scores might differ in claim, evidence and form from the individual student scores, and from group scores formed by aggregating individual student scores.

It was recommended that KDE and its contractors construct interpretive/use arguments as part of a validity argument for each reported score, and carefully evaluate whether scores can meet their intended interpretations and uses. In addition, KTAC members recommended careful consideration of possible conflicts or confusion arising from having scores and reports based on different evidence and interpretive/use arguments. For example, individual student performance is reported in terms of proficiency levels (NAPD) - would matrix-sampled group scores be reported in terms of NAPD distributions as well? Under what circumstances might it be possible for there to be different results of school NAPD performance based on aggregation of individual student results versus school results including matrix-sampled items?

Similarly, how comparable should student group and subdomain scores be in the multiple possible reports? KTAC members recommended that studies of the different reporting options be conducted in conjunction with other studies, especially scaling. Best practice in developing reporting scales, scores and reports would involve gathering input from users, e.g., through focus groups. They recommended that the studies' results be brought back for discussion before finalizing operational plans.

KTAC members discussed several possible options for creating scales for reading and the mathematics assessments at each grade. The scales should support the intended reports and uses. In addition, possible challenges were discussed, especially how to deal with potential multidimensionality at the level of subdomains, and between individual and group scores.

It was recommended that quality checks be considered for when individual and group scores might differ (e.g., consider when differential item functioning and differential test functioning analyses would be appropriate). KTAC members also recommended that scaling options include not only Rasch (1PL), but also 2PL (2-parameter logistic model), which might fit more complex skills better than the 1PL, 3PL (to account for guessing on multiple choice items), multi-level regression, bifactor, multi-level multi-dimensional regression and Bayesian models.

A key decision will be whether there will be one scale or multiple scales per grade/content area. Another consideration for multi-level models will be specifying and justifying what is fixed and what is random. KTAC members recommended that conceptual evidence (e.g., domain theory) and empirical evidence should be considered in evaluating the scaling options ("what works"), as well as communication ("will it be usable, acceptable, credible"). It was recommended that KDE and its contractor conduct additional studies, including simulation studies with appropriate data, and come back with recommendations and supporting materials.

Agenda Item: Science Assessment for 2022

Summary of Discussion: KDE and its contractor reviewed plans for the science assessment in 2022, including the test blueprint, administration, scaling and reporting. KTAC discussed the blueprint for 2022, which is a proportional reduction in content from the test blueprint used in 2019, when the science assessment's scale and performance level cutscores were established.

KTAC members recommended that validation studies be conducted to provide evidence that the items adequately measure the intended complex science constructs, especially at the lower end of the scale. It was recommended that KDE and its contractor ensure the "proportional reduction" in items by content area is captured in updated, specific test specification and blueprints and Performance Level Descriptors (PLDs), and check to ensure the new test blueprints support the intended interpretations and uses. The intended claims, as embodied in the PLDs, should be supportable, especially through analysis of alignment.

KTAC members also recommended that the reliability/precision accuracy and consistency of the new assessment be checked, both in terms of conditional standard errors across the score scale, and in terms of whether performance level decision accuracy and consistency are acceptable for accountability and other intended uses.

The reporting scale for science beginning in 2022 is desired to be consistent with the reporting scales that will be established for reading and mathematics. KTAC members recommended that the science scale be evaluated for appropriate resolution (e.g., no gaps in possible scale scores; small changes in scale scores are likely not to be interpreted as large differences in performance). Given the new reporting scale, KTAC members recommended that KDE and its contractor consider establishing a new calibration and a new underlying theta scale.

Agenda Item: Standard Setting for 2022 Reading, Math and Science Assessments

Summary of Discussion: Because new assessments will be administered in 2022 for reading, math, science, social studies and writing, new performance level cutscores need to be established for each grade/content area assessment. The discussion focused on standard setting for reading, writing and math. The same principles may be applied to social studies.

Standard setting for science was discussed separately because performance standards had been set previously for the science assessment in 2019. Kentucky has no requirements that the performance standards set for reading, writing or math be comparable to those of the assessments administered in 2019. With revised content standards and different assessments, KTAC members agreed that it made sense to use new reporting scales starting with 2022 results, and to instruct users not to compare results from 2022 and following years to results from 2019 and previous years.

KTAC members recommended that care be exercised in designing the standard-setting procedures to support appropriate decisions without incurring undue fatigue among the standard-setting panelists. Specifically, in the Bookmark standard-setting method, the construction of the Ordered Item Booklet (OIB) should provide items that are close together in the theta scale at areas around the cutscores.

Some ways to accomplish this are to: identify likely ranges for cutscores informed by policy or other information; using an adaptive (multi-phase) standard setting where ranges are identified using one OIB, and then using a finer-grained OIB for determining the final cut scores; and analyzing possible gaps in the difficulty of test forms items by analyzing the items used during the test construction process.

If the OIB is long enough that there is a threat of panelist fatigue, then some items might be culled, informed by both psychometric and content judgments. KTAC members recommended that the purpose(s) for using impact data be clear, and specific instructions be provided for how to interpret and not interpret impact data. Policy judgments should be either explicitly identified or avoided by the content-based standard-setting group (and perhaps addressed by a separate

group). This caution is especially applicable when 2022 impact data may be difficult to interpret due to reduced participation or lingering effects of interrupted learning.

KTAC members recommended that standard setting for writing follow the conception of the domain and the structure of the scales, since writing consists of separate scales for editing and mechanics, and for direct writing. It was recommended that KDE consider keeping the scales separate, and then combine for an overall score using a profile-based approach. Standard setting might be based on profile combinations of the two dimensions or also might also be based on a body of work to get the initial profiles, and then check with profiles of traits used in scoring. KTAC invited KDE to consider getting more generalizable and reliable information on writing by administering more than one writing prompt across the school.

It was recommended that for science standard setting, the Performance Level Descriptors (PLDs) that were developed and used to set standards in 2019 be carefully reviewed and edited if warranted to reflect both the desired claims and what can be supported by the new assessment blueprints. Statements regarding comprehensiveness, inference, and generalization should be carefully crafted (e.g., "student typically can do these types of things …" rather than "student can do [all] of these listed …").

Standard-setting instructions should match the intended claims. If PLDs are modified from the 2019 version, Kentucky content specialists should take the lead, supplemented with contractor expertise. The standard-setting plan should clearly identify whether PLDs will be adapted by the standard-setting panel, and if so, what instructions and guidance will be provided (e.g., to create borderline descriptors).

If the contractor uses previously existing PLDs and cutscores as the basis for conducting a standards validation in science, the contractor should show how the relevant content claims have not changed and how the performance standards established in 2019 have been mapped to the OIB used in 2022. The contractor should recommend particular standard-setting procedures based on their professional judgment regarding virtual versus in person, and if virtual, whether synchronous or asynchronous, etc.

The standard-setting plan should describe the threats and reasons for each recommended procedure (e.g., increases panelist motivation, is more efficient without loss of efficacy).

Agenda Item: Validity, Reliability and Impact Evidence for Kentucky Assessment Program

Summary of Discussion: HumRRO described the past and current work in which it has been engaged by KDE to help establish the validity, reliability and impact of Kentucky's assessment program. HumRRO also described the planning process for establishing future special studies, and invited KTAC input.

KTAC members commended KDE on having a long-running, systematic, extensive and credible program for gathering evidence of validity, reliability and usefulness, and for using such information to improve the state's assessment and accountability programs. KTAC members offered many suggestions of possible studies, acknowledging that KDE will need to prioritize what should be done within existing budgetary, available data and other constraints. Suggestions included:

- Reframe the approach to generating the research agenda: identify validity argument and associated claims (e.g., individual, group scores); identify the evidence needed to support, as well as critical threats (e.g., response process evidence for new format items intended to

assess higher order thinking skills); and prioritize studies to address what is needed for the validity argument (including use and impact).

- Evaluate the quality and usefulness of new group level scores – do they work? How well? Attend particularly to the validity of interpretation and communication (e.g., how to avoid an ecological fallacy – more homogeneity interpreted between schools than exists, where students fall around group score point). This will entail considering the relation to school-level precision and student groups.
- Now is a very good time to monitor and evaluate how well the test design supports the intended uses (e.g., interpretation, claims, test design in science).
- Include evaluations of fairness (e.g., COVID-19 differential learning opportunities and long-term effects on student learning and performance), and monitor for potential effects on scaling designs, standard setting and reporting by student groups. It should be pursued how to use information gathered on OTL - and whether the most productive OTL information was being collected - especially for points of key impact or transitions (e.g., grade 9 as a gateway grade). It was noted that HumRRO includes studies of fairness under other legislative categories.
- Consider how standards and performance link to the real world (e.g., what can be learned from the Atlas-DLM approach and other work specifying college/career competencies), especially for ACT and CTE assessments that are close to postsecondary outcomes.
- Understand ACT alignment with standards, and performance correspondence with other (middle school) tests.
- Investigate how assessment data are used (e.g., to inform instruction) Beyond descriptive studies, evidence of consequential use and impact would be very valuable. This might start with evaluations of the design of assessment reports and the effectiveness of training to interpret reports.
- As the local assessment and accountability programs get underway, support should be provided to gather validation evidence in design, implementation and impact.
- Evaluation should continue of the accountability program - the suitability of the indicators (e.g., achievement, improvement, school climate and safety) and how they are measured, and in particular how the accountability design influences school scores, ratings and identification, and what actually fuels student learning and school improvement.

Agenda Item: Local Innovative Assessment and Accountability Development

Summary of Discussion: KDE provided an overview of the current work sponsored by the department under its advancing education initiative, including fostering innovations in assessment and accountability design and use through the work of the Kentucky Coalition on Advancing Education.

KTAC members noted that the field of measurement acknowledges now more than ever that it is important to be sensitive to what constructs are valued, how they are defined and who gets to define them. This is essential for validity and equity. KTAC members complimented Kentucky on this awareness being built into support of local efforts to develop innovative assessment and accountability systems.

They encouraged KDE to provide these local initiatives with guidance and tools early in the process to gather and effectively use ongoing program evaluation information. Such documentation will be essential to validate the appropriateness of decisions, such as what was chosen to focus on, and how the assessment designs were decided upon. Ongoing program evaluation also will also be essential in helping the projects address challenges while they can be more easily solved, and in documenting more completely lessons learned to help other users.

Agenda Item: Future KTAC Meeting Dates

Summary of Discussion: KDE staff and KTAC members agreed it would be beneficial for the KTAC to meet to review and advise on work done to further develop Kentucky's assessments, especially in preparation for the 2022 administration. The following dates and times were agreed upon (all times Eastern):

- Feb. 4, 2022 2:30-5 p.m.
- Mar. 14, 2022 2:30-5 p.m.
- June 22, 2022 2:30-5 p.m.
- Oct. 6, 2022 12-5 p.m.

At this time, all meetings will be held virtually via Microsoft Teams. There was discussion on extending the Oct. 6 date to make it an in-person meeting. However, that will be determined later.

The meeting adjourned at 5:30 p.m. ET.